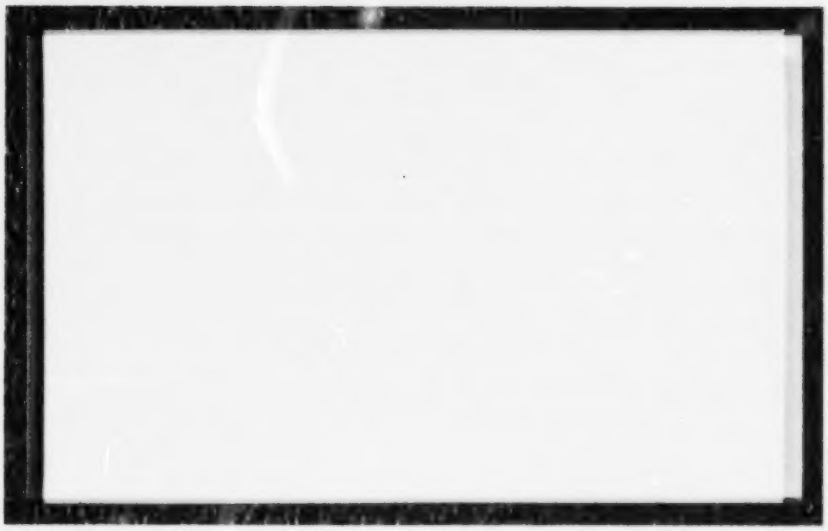
UNIVERSITÉ DE MONTRÉAL



CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

CRT-590

AN OPERATIONS RESEARCH APPROACH FOR THE PLANNING AND OPERATION OF AN AMBULANCE SERVICE

by

Pierre Trudeau, Jean-Marc Rousseau,
Jacques A. Ferland, Jean Choquette

Cette étude a été publiée grâce à une subvention du fonds F.C.A.R. pour l'aide et le soutien à la recherche.

Université de Montréal
Centre de recherche sur les transports - Publication #590
Août 1988

Dépôt légal - 3e trimestre 1988

ABSTRACT

Emergency medical services are very complex systems to operate. Most of the subproblems are large scale and many are stochastic in nature. After describing the problem, we present the results obtained in four principal areas: demand forecasting, scheduling, optimal location of waiting points and simulation as an evaluation tool.

Key words: EMS; demand forecasting; scheduling; location; simulation.

- - - - - - - - - - -

RÉSUMÉ

La majorité des centres urbains sont aujourd'hui dotés de systèmes de véhicules d'urgence. L'exploitation de tels systèmes suscite des problèmes considérables, notamment en raison de la taille des problèmes et de la nature aléatoire de plusieurs des processus en jeu. Après une description de la problématique en jeu, nous présentons les résultats obtenus dans quatre axes principaux : la prévision de la demande, la confection des horaires du personnel, la localisation optimale des postes d'attente et la simulation comme outil d'évaluation.

Mots-clés : service ambulancier; prévision de la demande; horaires; localisation; simulation.

1. INTRODUCTION

Most urban communities currently have their own emergency medical services (EMS). However, the operation of such systems is very complex. Of course, similar complex problems also arise in other contexts, but here the quality of service has a direct impact on the health of the population, and the gain is expressed, amongst other things, in terms of saved lives.

The objectives of an EMS are primarily specified in terms of medical considerations, but financial and legal factors also greatly influence the formulation of the problem. Efficiency and equity are the main preoccupations of the authorities. The satisfaction of demand within a specified response time is the commonly expressed objective. But the terms "satisfaction", "demand" and "response time" are often vague and full of ambiguity. Their precise meaning and interpretation will greatly impact the level of resources needed. Once objectives have been clearly defined, the problem boils down to determining the quantity of resources required at every time of the day, and how should they be used in order to fulfill the objectives.

This problem has many facets that can be briefly summarized as follows:

i) Demand forecasting: demand is stochastic both in space and in time. Since the level of resources in service should vary accordingly, demand forecasting becomes an important input for the scheduling of units and their personnel.

ii) Demand conversion: after computation of forecasts, demand has to be converted in number of vehicles to be in service at all times.

iii) Scheduling of personnel: this includes not only personnel for the units but also personnel answering calls. The latter must be adjusted to demand in order to minimize occasions when all

lines are busy. Staff answering calls must decide on the type of service required before a request is transferred to the dispatching unit. Their medical qualifications influence their capacity to filter calls.

iv) Operational strategy: an operational strategy involves decisions concerning:

   a) location: where to localize the waiting points;
   b) dispatching: which free unit should answer the call;
   c) relocation: when and how should the free units be relocated to the waiting locations;
   d) anticipation: how can the system be organized to prepare for future events which occur either with certainty or with specified probability.

v) Minimization of internal and external delays: here, we group together potential sources of time loss, for which reductions often depend on technological tools allowing information to be transmitted more efficiently. Delays may occur either inside the operation center (internal delay) or outside (external delay). They include the time required to establish a diagnosis, the transfer of data between the answering personnel and the dispatchers, the localization of free units, the identification by the units of shortest time paths and the choice of the hospital (offering appropriate treatment with an emergency room not too overloaded).

vi) Personnel issues: the qualifications of personnel should be adequate at each stage of the overall process to ensure appropriate handling of calls for service. This is particularly

true at the two occasions when personnel is in contact with the patient: at the reception of the call (on the phone) and when a unit arrives on site. The proportion of MDs, paramedics, nurses, stretcher-bearers and non medically qualified operators used has a great impact on how an EMS reacts to a specific environment.

For the past three years, the authors have been active in developing a coherent and complete set of tools for planning and operating emergency vehicle systems. The collaboration of Urgences santé (the body in charge of the EMS in Montreal and suburbs) was essential for the success of our project. However, we did not limit ourselves to their context, but rather examined a broader diversity of environments.

The objective of this paper is to show how operations research can help in solving real world problems related to the planning and operation of an ambulance service. A selection of these problems is presented and on each occasion, the main lines of the proposed solution approach and computational results are presented.

The problem includes both quantitative and qualitative considerations, but in our study we are concerned only with the first. The first step in analyzing such a complex problem is to perform demand forecasting. Section 2 is devoted to this. From this analysis, we determine the amount of resources required at each time of day. The next step is to specify how these units and their personnel are to be put into operation. In Section 3 we deal with issues related to the scheduling of units and personnel. Analytical procedures for determining the location of waiting points are introduced in Section 4. Finally, in Section 5 we propose a simulation tool to evaluate the quality of any strategy for operating an EMS.

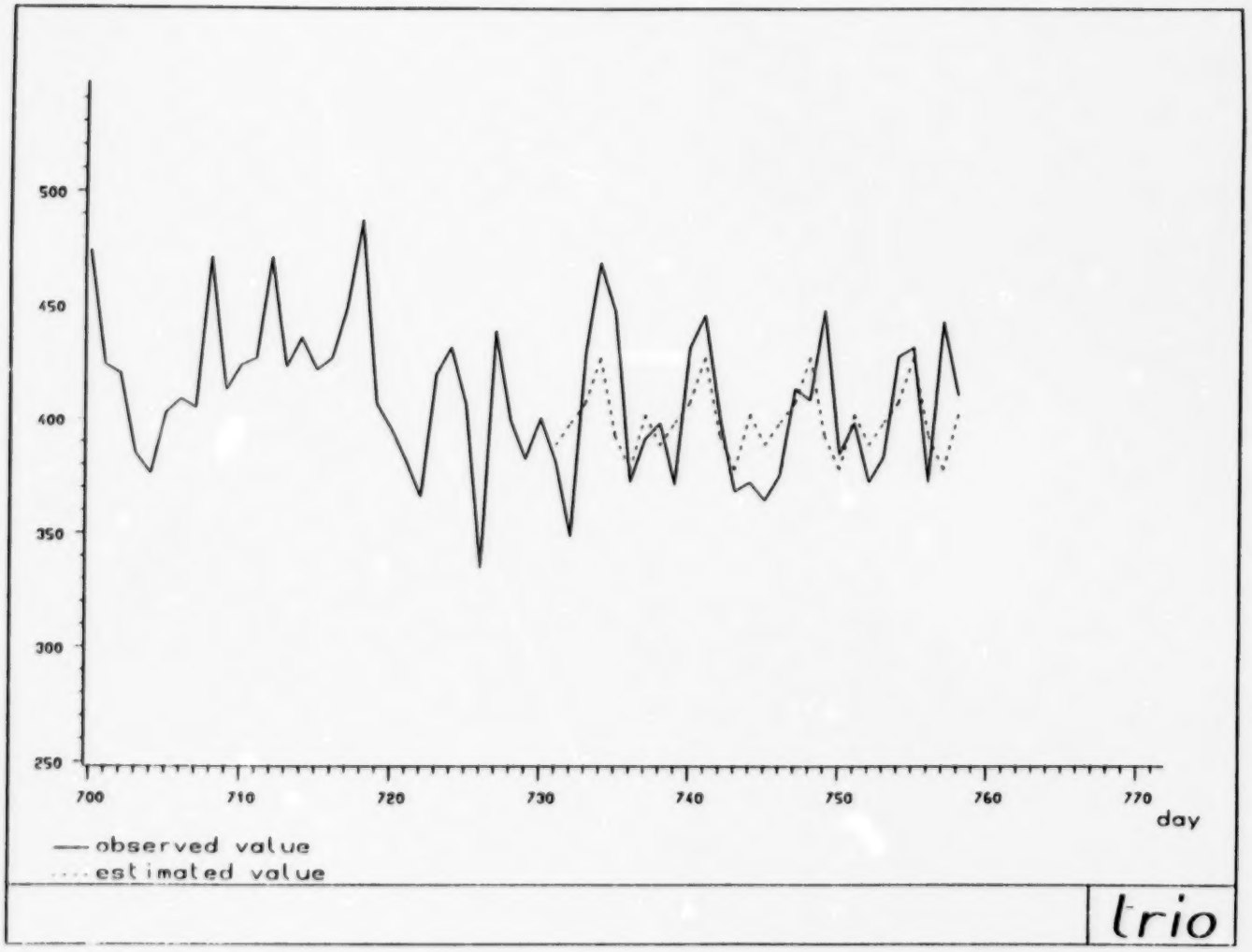


Figure 1: Daily forecasts obtained by Box-Jenkins method

The graphics presented in this paper were produced using TRIO (Babin et al., 1987), an open interactive graphics program developed at the Centre de recherche sur les transports in Montreal for the formulation and simulation of regression models.

2. DEMAND FORECASTING

The level of resources in service should vary according to the demand level. Since demand is stochastic, forecasting is essential to determine the amount of resources required and to prepare vehicle and crew schedules.

In Montreal, the amount of resources in service traditionally varies from hour to hour. Hence forecasting should also be completed at the same level of detail. The problem is split into two parts: a forecasting model for the daily level of demand, and another one for the hourly level. Combining both models, we disaggregate the daily forecasts into hourly forecasts.

2.1 Daily models

Amongst others, ARIMA models and econometric models based on regression techniques can be used to forecast daily demand. ARIMA models are autoprojective. A forecast Y_t depends only on past observations $Y_1, \ldots, Y_{t-1}$, and on a random perturbation u_t, i.e., $Y_t = f(Y_1, \ldots, Y_{t-1}, u_t)$. Using a relatively small number of parameters one can obtain a wide range of models. Identification procedures enable us to select the one with the best fit. We can then estimate the parameters and compute optimal forecasts (Box and Jenkins, 1976; Gouriéroux and Montfort, 1983).

In contrast, in an econometric model, the random variable Y_t is a function of a vector of explicative variables X_t, called exogenous variables, and of a random perturbation u_t, i.e., $Y_t = f(X_t, u_t)$. X_t can be

deterministic or stochastic and independent of u_t. If the u_t are correlated or if X_t contains past values of Y_t (then called lagged endogeneous variables) then the econometric model is said to be dynamic. For application of these models to transportation see Dagenais et al. (1985) and Gaudry (1984).

Several variants of these two classes were used to compare both approaches. In order to do this, Urgences santé supplied us with data from September 1st 1984 to December 31st 1986 (851 observations overall). Part of the data (the first 730 observations) was used for building the model and the other part for testing. Figure 1 shows the results obtained using the best Box-Jenkins model over a horizon of 28 days (day 731 to day 760). For clarity, only the observations from day 700 onward are shown in Figure 1. The results (table 1) indicate that both approaches are equivalent, the best ARIMA model being sligthly better than the best econometric model. Furthermore, in all cases the Forecasting Mean Square Error was below 10% (Choquette et al., 1988).

Table 1
Best Daily Forescasting Models

Model type	Model equation	Relative absolute error	Mean square error
ARIMA	$(1-\phi b)(1-b^7)\ Y_t = (1-\theta b)(1-\alpha b^7)u_t$	0.05	27.27
Regression	$Y_t = \beta_o + \sum_{j=1}^{6} \beta_j I_{jt} + u_t$	0.05	27.48

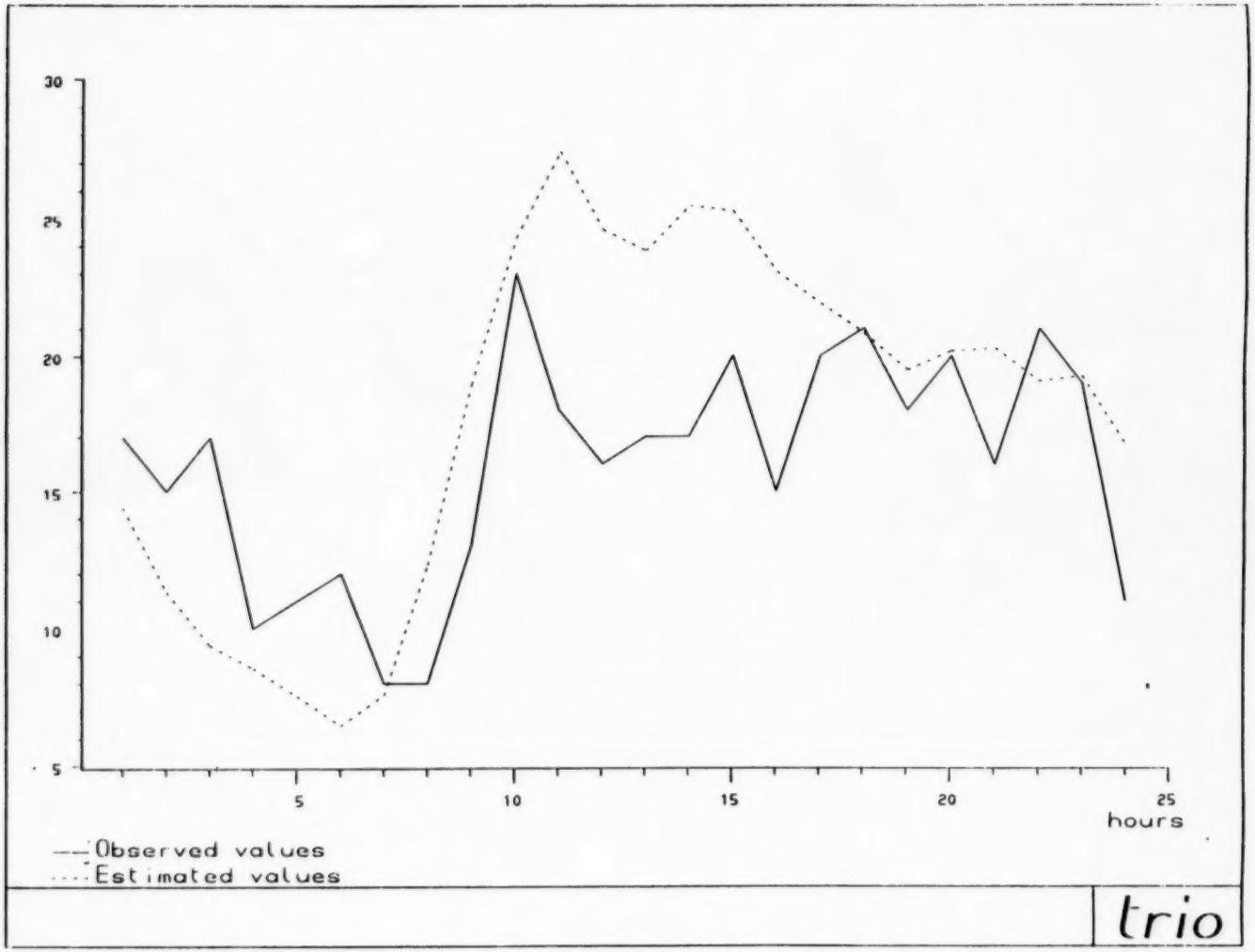


Figure 2: Forecasts of the Hourly Number of Calls

2.2 Hourly model

Preliminary studies suggested that the hourly distribution of demand varies with the day of the week. Furthermore, it may also be dependent on seasonal changes. In our hourly model, the forecasted values must be between 0 and 1 (representing the proportion of daily demand occurring each hour), and care must be taken to deal with autocorrelations.

A market share model, called LOGIT, was used to disaggregate the data (Domenchich and McFadden, 1975). To reduce the computer memory space required, each period is three hours long. The LOGIT model applied in this context determines the proportion of calls during each period as follows:

$$p(i) = \frac{e^{V_i}}{\sum_{j=1}^{8} e^{V_j}}$$

where p(i): proportion of daily calls during period i;

V_i : utility function associated with period i. In our case V_i is a linear function of seasonal-S (harmonic variables), daily-D (dummy variables for days of the week and holidays), and hourly-H (climatic conditions) variations:

$$V_i = \beta_{io} + \sum_{k=1}^{n_S} \beta_{ik} S_k + \sum_{m=1}^{n_D} \beta_{im} D_m + \sum_{n=1}^{n_H} \beta_{in} H_n \ , \quad i=1,\ldots,8$$

Figure 2 illustrates the forecasts obtained at the hourly level by the two stages procedure for one day.

2.2 Conclusion

For both the daily and hourly models, the best models turned out to be the simplest models. Very few of the large number of parameters first considered remain in the regression model, as shown by the model equation in table 1. Instead of proceeding in two stages, an ARIMA model that builds directly from the hourly data is presently under study.

The number of vehicles to be in service at all time should be extracted from the forecasted demand by day and hour. This conversion is by itself a complex process that must take into account the objectives of the EMS, the operational strategy adopted, the service time distribution of calls, the variability of the demand that makes the system in a perpetual transient state, etc. This facet of the problem was not addressed in our project.

The next section focuses on schedules and it is assumed that the demand expressed in numbers of vehicles has already been determined.

3. SCHEDULES FOR AMBULANCES AND AMBULANCE DRIVERS

It costs about half a million dollars in Montreal to operate a unit each hour of the year. Needless to say, scheduling should be carried out to fit the demand as exactly as possible, without using more units than necessary. To reflect variations in demand, the Montreal EMS determines a different schedule for each season and this schedule is cyclic with a period of one week.

The demand forecasting process should therefore identify a typical week for each season. From now on, we assume that such a typical week has been identified and that the demand is known and expressed in number of units for each hour of this typical week.

In our model, the typical week is partitioned into four blocks: M-T (Monday to Thursday), F (Friday), Sa (Saturday) and Su (Sunday). Hence a typical day is used to describe the demand of each block.

The design of the various schedules is constrained from many sources (physical number of units, telecommunication facilities, ...) but many of these originate within the collective agreement between employees and the authorities.

The following definitions are used to specify the constraints, and a list of the most important constraints follows.

- Definition

  piece of work : a continuous period of time during which a team of unit employees operates the same vehicle;

  unit duty : a daily assignment consisting of one or two pieces of work representing the total use of the vehicle for the day;

  circular schedule: refers to the interpretation of the demand of a typical day as repeating itself day after day. For example, a piece of work starting at 22h00 with length of 10 hours will cover the hours 22, 23, 24, 1, 2, ..., and 7 of this same typical day.

- List of constraints

  C1 - The pieces of work must be 8 or 10 hours long.

  C2 - Between any two consecutive pieces of work for the same vehicle, there must be a pause of at least one hour (for maintenance).

  C3 - The pieces of work must start during the period from 6h00 to 24h00.

  C4 - The number of radio channels dedicated to communications between dispatchers and unit personnel is limited to three. Since the team must report at the beginning of each piece of work, the number of simultaneous starts is limited to three and there must be five minutes between two consecutive starting times.

  C5 - The total number of units is limited.

C6 - Since the fleet of vehicles is owned by several private companies, the workload must be assigned according to the proportion of vehicles they own.

C7 - Each ambulance employee must work 80 hours per two weeks. It is tolerated that they work between 76 and 84 hours. They are also entitled to one week-end off every two weeks.

We now present the general algorithm producing the unit duties, their assignment to the companies, and finally the employee schedules. The solution techniques are then briefly discussed.

STEP 1: Generation of pieces of work

a) Solve circular schedule problems for blocks M-T and Sa.

b) Solve non-circular schedule problems for blocks F and Su, taking into account the effect of M-T schedule on Friday morning and the effect of Sa schedule on Sunday morning.

STEP 2: Shifting of the pieces of work and generation of unit duties

a) Shift the pieces of work five minutes apart.

b) Generate unit duties.

STEP 3: Unit duty assignment to ambulance companies

- respecting equity
- trying to obtain homogeneous schedules.

STEP 4: Ambulance personnel schedules.

3.2 Generation of pieces of work

Aggregating steps 1 and 2 and expressing the demand and the departure time in units of five minutes would generate an integer program that is too large. In step 1 the unit of time is one hour. It is then possible to enumerate all possible unit workdays respecting constraints C1, C2, C3 and C5. Constraint C4 is aggregated at the one hour level. Then,

using integer programming we find the number of each workday type that should be used to minimize costs. There are three types of cost: i) one for overcover, ii) one for undercover (which is higher than the previous one) and iii) one for hourly salary. The structure is such that i) and ii) dominate iii) allowing a choice between two equivalent solutions in terms of i) and ii), the one with the least number of hours worked.

We use here the HASTUS-macro code (Blais and Rousseau, 1981) developed in the context of urban transit bus scheduling problems.

3.3 Shifting the pieces of work and generation of unit duties

Since each unit starts at the beginning of an hour, the solution in step 1 does not necessarily respect constraint C4. In order to do this we shift the departures but only by five minutes from the scheduled hour. However, these shifts delay the time at which the required level of coverage is attained. In periods of increasing demand, for instance, it is desirable that the additional units required, over and above the previous period, be in service as early as possible.

This leads us to the notion of "perfect" and "imperfect" cycles. A perfect cycle is a set of three (in our case) pieces of work such that the end of one piece of work coincides with the beginning of another. For instance, the three pieces of work 8h00-16h00, 16h00-24h00, 0h00-8h00 together form a perfect cycle. An imperfect cycle is identical to a perfect cycle except that the loop is not completed to cover the whole 24 hours. For instance the two pieces of work 10h00-20h00, 20h00-4h00 together form an imperfect cycle because the interval 4h00-10h00 is not covered.

Perfect cycles are interesting, for if we synchronize the shift of the pieces of work that constitute them, then we preserve one ambulance coverage all day, without being penalized. It becomes advantageous to first shift the imperfect cycles to the first five minutes of the hour and the perfect ones afterwards.

If constraint C4 remains violated, the cycles including a piece of work scheduled at a five minute point violating it are broken and their pieces assigned individually.

The original ambulance days of step 1 have been perturbed in step 2a. Therefore, instances where constraint C2 is violated may have been created. In step 2b we attempt to minimize if not eliminate such cases. The problem is complicated by the fact that in Montreal constraint C5 is strong.

An initial assignment is first obtained solving a matching problem using the solution of step 2a without any modifications. Then, an improvement phase takes place where we shift, first sequentially then simultaneously, the cycles involving the unit duties not respecting constraint C2.

3.4 Unit duty assignment to ambulance companies

The four blocks M-T, F, Sa and Su are processed sequentially.

1) For the first block, a matching is performed with the duties of a previous schedule in which the equity objectives were almost perfectly fulfilled. The matching costs are a function of the starting hour and duration of the piece(s) of work of the unit duties.

2) Exchanges of duties are performed between companies to improve equity.

3) Steps 1) and 2) are repeated for the three other blocks. In 1) we use the cumulative result of the blocks already treated instead of a previous schedule.

4) For every company the pieces of work of the M-T and F blocks, are combined to obtain the potential number of crews departing each hour. The same is done separately for Sa and Su blocks.

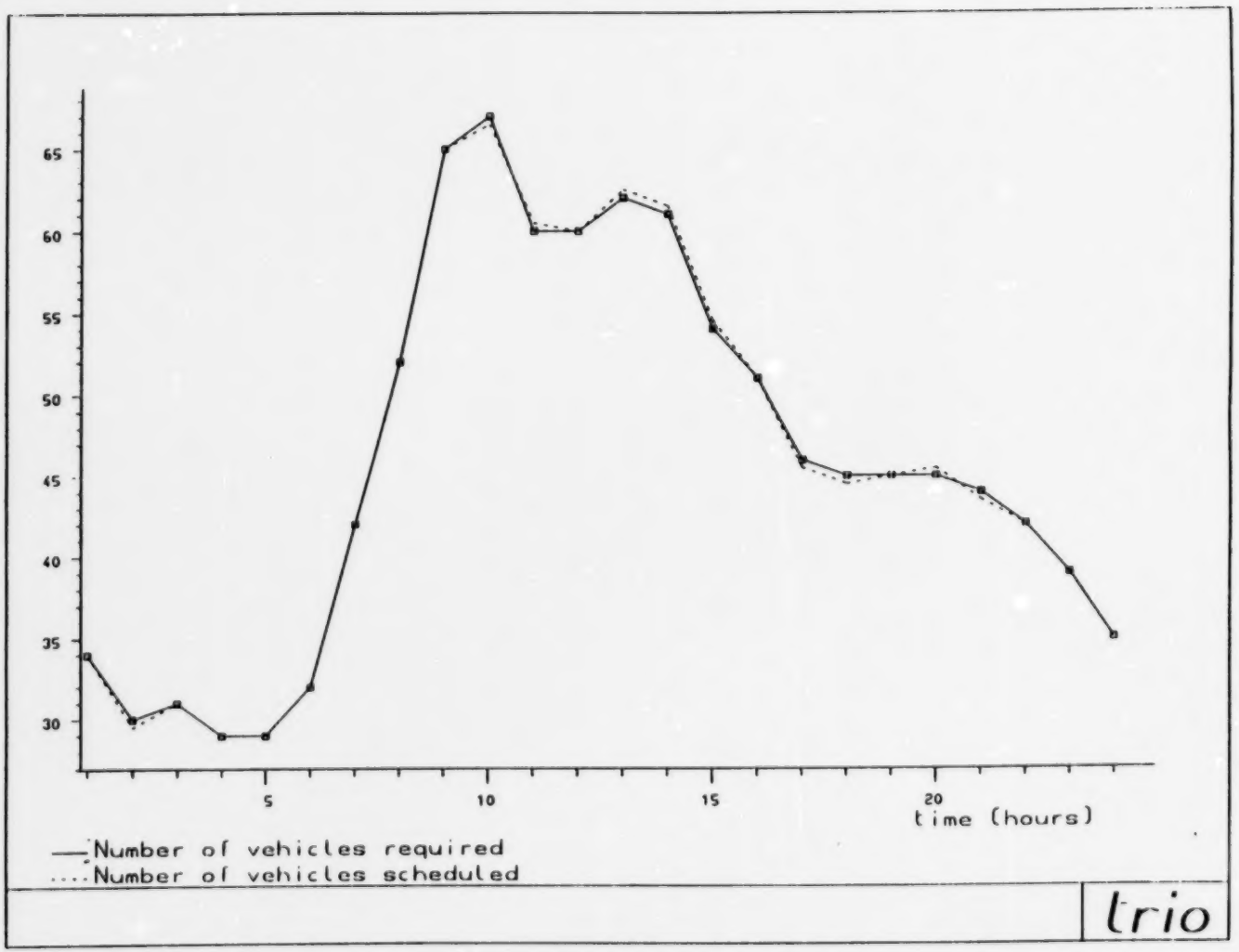


Figure 3: Weekly schedule (Monday-Thursday) - Summer 87

5) A pool of pieces of work is formed by the excess weekend duties over the M-F duties. These are then reassigned to the companies having a deficit during the weekends.

6) At the end, a few pieces of work remain. These are carried out by part time employees.

3.5 Ambulance personnel schedules

Each company now has its set of pieces of work. The companies are processed sequentially. In order to deal with constraint C7, we use a two week horizon, duplicating the pieces of work of the typical week.

1) Only a very limited number of acceptable 14 day patterns are considered first. A pattern is composed of 10 binary values (for the weekdays) indicating if the crew is working or not that day. One weekend is composed of two holidays and the other of two working days.

2) We first rigorously require that all pieces of work of a pattern start within the same hour. This is then progressively relaxed. We examine each pattern type in a rolling sequence, filling it when possible.

3) A matching between Sa and Su pieces of work is performed where the costs are a function of the differences in departure times. The resulting pairs are labelled with their average departure times. The valid patterns are also labelled with their average departure times.

4) Steps 2) and 3) are first executed using only 8 hours duration, then only 10 hours duration and finally with a mix of the remaining pieces of work. Then new valid patterns are introduced as required.

Overall, more than 95% of the pieces of work are within a valid schedule. The rest will be executed by part time employees.

3.6 Conclusion

The algorithms that we have described have been in operation for two years now with the Montreal EMS. They are producing better schedules, much more rapidly, showing great flexibility. The schedules were previously produced manually and the number of vehicles scheduled could be up to four units above or under the number required. Figure 3 shows how well the demand curve can be fitted now, respecting at the same time all the side constraints coming from the collective agreement, telecommunication limitations, etc. Note that halves of vehicles can be obtained in the schedule because each piece of work is entitled to a thirty minute break.

In figure 3, the absolute difference between the number required and the number scheduled is at most 0.5. The whole process of vehicle and crew scheduling uses about eight hours of CPU time (on a PC-AT microcomputer). This rapidity is of great value when one needs to evaluate the impact of modifying the value of various parameters, notably in the context of collective agreement negociation. Given the number of vehicles at each period of the day we must then locate them on the network and dispatch them. We now consider this problem.

4. OPTIMAL LOCATIONS OF WAITING POINTS, RELOCATION OF VEHICLES AND DISPATCHING RULE

The location of vehicles waiting to be dispatched, their relocation, and the dispatching rule have a direct impact on the operational procedures and their efficiency to respond to demand. Other stragegic issues related to anticipated occurrences of specific events may also have bearing on the operations but were not treated in our study.

Of the large set of dispatching rules that one can imagine, the most natural rule is certainly "the closest available unit answers the call". Although Carter et al. (1972) showed that this rule is not optimal (the

optimal policy is generally state dependant), it is quite acceptable since Jarvis's (1981) results indicate that using the optimal policy induces an improvement of less than 2 % over the solution obtained with this simple rule. We therefore adopted this dispatching rule.

The location and relocation issues are quite intermixed. In the location problem, one seeks the sites of the waiting points i.e. the points occupied by the available units. On the other hand, relocating means reassigning currently available units to new waiting points. It may be necessary for example, to fill a gap that may have developed because of the concentration of calls in a specific region.

We have classified the possible strategies related to location and relocation as follows:

Strategy A: no relocation allowed.

Each ambulance station has a specific location and has to respond to calls coming from a specific district. Furthermore each station has its own units that always return to that station after servicing a call.

case i): no cooperation.
A call from a district can only be handled by a unit from the station responsable for the district. If no unit is available, then service is delayed.

case ii): cooperation allowed.
In this case, if no unit is available at the district station, a unit from a neighbouring district can handle the call.

Strategy B: relocation allowed.

case i): with fixed waiting points.

Units are not specifically linked to a station and they can potentially occupy any of the waiting points.

case ii): patrol.

Each free unit is in movement patrolling a given territory which can be modified as the number of available resources varies. Allowing more freedom, this situation, though common for a police department, is rarely seen in EMS.

In fact, starting with these four basic strategies one can imagine almost an infinite number of variants. In Montreal, for instance, a variant of strategy B i) is used where 25 specified waiting points have to be covered permanently using relocation if necessary.

Analytically, we first looked at the problem of identifying optimal locations under strategy A ii). Larson and Jarvis made important contributions in this field of research. They dealt explicitly with the stochastic nature of the system where in particular the servers (the vehicles) are not always free to answer calls. This model is called the hypercube model (Larson, 1974; Larson and Odoni, 1981). The region to be covered is partitioned into sub-regions, each represented by a node. Each free ambulance can be positioned on one or many nodes (simulating patrol in the latter case). The location of waiting points are fixed and no relocation is allowed.

The demand is assumed to occur according to a Poisson process with constant parameters, each node being an independent source of call. Each call is assumed to require only one unit for service. The dispatching rule has to be specified a priori but can be of any nature. Furthermore when all servers are busy, one can drop the call, use a special resource assumed to be always available, or build a queue managed according to a policy independent of the call's origin. The service time of each vehicle is assumed to be exponentially distributed and independent of the server, of the call origin and of the past history of the system.

Each state of the model is represented by an ordered set of cardinality N (N being the number of vehicles), each component indicating if the corresponding vehicle is busy or free. When no queue is allowed, there are 2^N states, corresponding to the corners of a unitary hypercube in R^N. One obtains a continuous time discrete state Markov process. Larson (1974) proposes a method to compute exactly the steady state probabilities for a system including up to about 12 vehicles.

In Larson (1975) an approximate procedure is proposed to deal with large problems where a system of N (nonlinear) equations has to be solved rather than one including 2^N (linear) equations. The error induced by this procedure is reported to be less than 2 %.

Note that the hypercube model is only a descriptive tool generating performance statistics for a specified strategy. Jarvis (1975) proposed to embed the hypercube in an iterative process to seek an optimal solution. "Sending the nearest available unit" is the dispatching rule adopted. The algorithm is summarized as follows:

Step 1 - Define an initial solution, specifying the location of the N vehicles

Repeat

Step 2 - According to "sending the nearest available unit" dispatching rule, determine the dispatching table D where D[n,a] specifies the a^{th} preferred unit (i.e. when all the a-1 first preferred units are busy) to answer calls originating at node n;

Step 3 - According to this dispatching table, determine the optimal location of the units.

Until (solution does not change) OR (deterioration of the objective).

The Jarvis procedure converges but does not guarantee an optimal solution. For small problems Jarvis reports good behavior of the procedure to converge to a global optimum. However, implementing the same procedure on the Montreal network with 484 nodes and as many as 60 units, we found that the approach in Step 3 generates a strong inertia to remain close to a local optimum.

We propose a variant of this procedure where Step 3 is modified to analyze dynamically the influence of changing the location of only a few units at a time. The procedure is summarized as follows (the reader is referred to Trudeau and Hodgson 1988 for technical details). For simplicity we describe it for the case where only one unit is moved at a time.

- Modify the location of a unit to one of its neighbour point (e.g. one of the 10 closest points).
- Update partially (to reduce computations) the D table, keeping only a subset of lines (those "strongly" affected by the change) and a subset of columns (that increases with ρ the average workload of the system). The estimation of the performances of the new solution is based on this partially modified D matrix;
- If an improvement is induced, repeat the procedure with other neighbour points.
- The units are sequentially moved until a stable solution is reached.

Some numerical tests were conducted with this procedure. They all used the same network (484 nodes) and the same distribution of demand among the nodes. We now report on these experiments.

- We examined nine scenarios, each described by the pair (N,ρ), where N is the total number of ambulances available and ρ is the average workload of the system. N could take the values 10, 30 or 50 and ρ the values 0.2, 0.4 or 0.8.

Table 2.

Computational results

N	ρ		JARVIS Ave. R.T.	JARVIS CPU	DYNAMIC APPROACH Ave. R.T.	DYNAMIC APPROACH CPU	IMPROVEMENT (%)	GLOBAL IMPACT (%)
10	0.2	Worst	3m59	1 021	3m53	816	0.0	
		Best	3m48	581	3m44	124	5.9	
		Variation (%)	4.8		4.0			1.8
10	0.4	Worst	4m56	1 057	4m50	124	0.9	
		Best	4m52	584	4m47	501	2.4	
		Variation (%)	1.4		1.0			1.7
10	0.8	Worst	8m57	1 173	8m54	849	0.1	
		Best	8m53	452	8m52	234	0.9	
		Variation (%)	0.8		0.4			0.2
30	0.2	Worst	2m21	2 668	2m13	11 184	5.2	
		Best	2m15	1 755	2m05	5 517	8.4	
		Variation (%)	4.4		6.4			8.0
30	0.4	Worst	2m54	3 573	2m44	8 011	2.8	
		Best	2m45	2 267	2m39	5 396	7.8	
		Variation (%)	5.5		3.1			3.8
30	0.8	Worst	6m28	2 744	6m20	6 382	0.8	
		Best	6m22	1 357	6m19	4 233	2.4	
		Variation (%)	1.6		0.3			0.8
50	0.2	Worst	1m53	5 767	1m40	17 955	5.1	
		Best	1m43	4 220	1m38	11 854	11.6	
		Variation (%)	9.7		2.0			5.1
50	0.4	Worst	2m24	6 459	2m03	18 105	4.9	
		Best	2m09	2 881	2m02	12 047	14.9	
		Variation (%)	11.6		0.8			5.7
50	0.8	Worst	5m17	6 636	5m02	21 688	1.1	
		Best	5m05	2 944	5m01	18 306	5.0	
		Variation (%)	3.9		0.3			1.3

- For each scenario (N,ρ), five initial solutions were randomly generated.
- Then the Jarvis algorithm was applied until convergence was obtained.
- Our "dynamic procedure" was applied to Jarvis solution until convergence was obtained. This was done to reduce overall CPU time and to see how far the procedure could escape from local optimum solutions.
- The objective was to minimize the average response time. The results are summarized in Table 2.

As an example, the first three lines of Table 2 should be read as follows:

For 10 waiting points (N=10) and an average workload of 0.2 (ρ=0.2), among the five random initial solutions, the "worst" result for Jarvis was 3m59 and the "best" 3m48; the "worst" CPU time (not necessarily corresponding to the worst solution) was 1 021 sec. and the "best" 581 sec. The difference between the best and worst solutions produced a "variation" of 4.8% (3m59/3m48 = 1.048) in the average response time. The same interpretation holds for the dynamic approach. Note that the "worst" solution for the dynamic approach is not necessarily the result of applying this approach to the "worst" solution obtained using Jarvis approach. It is simply the worst solution of the five trials. In the next column, 0.0% was the "worst" improvement obtained by applying the dynamic approach to the Jarvis solution among the five trials and 5.9 the "best", whereas the last column shows the saving when we compare the "best" solution of each approach (3m48/3m44 = 1.018).

The following observations can be derived from these results:

- Jarvis solutions vary greatly between them for a given (N,ρ) and this is amplified as N increases to reach a 11.6 % variation (for N=50 and ρ=0.8). This confirms the problem of local optima.
- The dynamic approach allows escape from local optima and generates final solutions showing less variability.
- Although Table 2 does not demonstrate this, each of the five final solutions obtained for a given (N,ρ) pair by the dynamic approach is better than the best solution obtained using Jarvis approach (except for (N=10, ρ=0.2) and (N=10, ρ=0.8).
- The improvement is most significant for ρ=0.2 and ρ=0.4, which corresponds precisely to the practical situation in Montreal.
- The improvement seems to increase with N to reach up to 15%.
- The CPU time required by the dynamic approach is much larger than that for the static approach. Though there was no emphasis on efficient programming, it seems appropriate to consider a hybrid approach where the static approach could be used during the first few iterations.

The work presented in this section represents one of the theoretical sides of our project. It did not give rise to a direct application since contrary to conditions for the Montreal EMS, relocation of the units is not allowed. However, this type of study could give indications on optimal location of waiting points that could be embedded in an other strategy. In addition, we wanted to initiate the building of the largest possible range of strategies in order to better evaluate the relative merits of each of them, in particular the one currently implemented in Montreal. However, before implementing any new location and dispatching strategy, it is essential to thoroughly evaluate it. Simulation is the best tool for this.

5. SIMULATION

The complexity of the framework of any EMS requires flexible tools allowing the user to evaluate several strategies and to identify the one which seems the "best" according to specified objectives. The need for such tools is critical in the EMS context since experiments should be avoided because lifes could be put in jeopardy. Furthermore, even if important developments have been completed on analytical models during the last fifteen years, some strong hypotheses (exponential service time and constant arrival rate) rarely verified in practice are often required. Also, relocation strategy can be analyzed only in very restrictive contexts (Wesolowsky and Truscott, 1975; Kolesar and Walker, 1973; Berman, 1981a, 1981b).

For these reasons a simulation model has been developed to verify the likelihood of results obtained with analytical models. This model cannot generate an optimal solution but can evaluate any strategy of interest.

Although it was developed in collaboration with the Montreal EMS, our simulation model remains as general as possible. With slight modifications it can deal with other emergency systems such as police or fire departments. This discrete-event simulation model (Bratley et al., 1987) is of a finite horizon since the demand in a real EMS is not stationary. Infinite horizon simulations are performed exceptionally when, for instance, one wants to gradually perturb the hypothesis of analytical models and measure their impact on steady state performances.

5.1 States, events and decisions

The state of the system is the status of each component entering into the model. In our model, these components are:

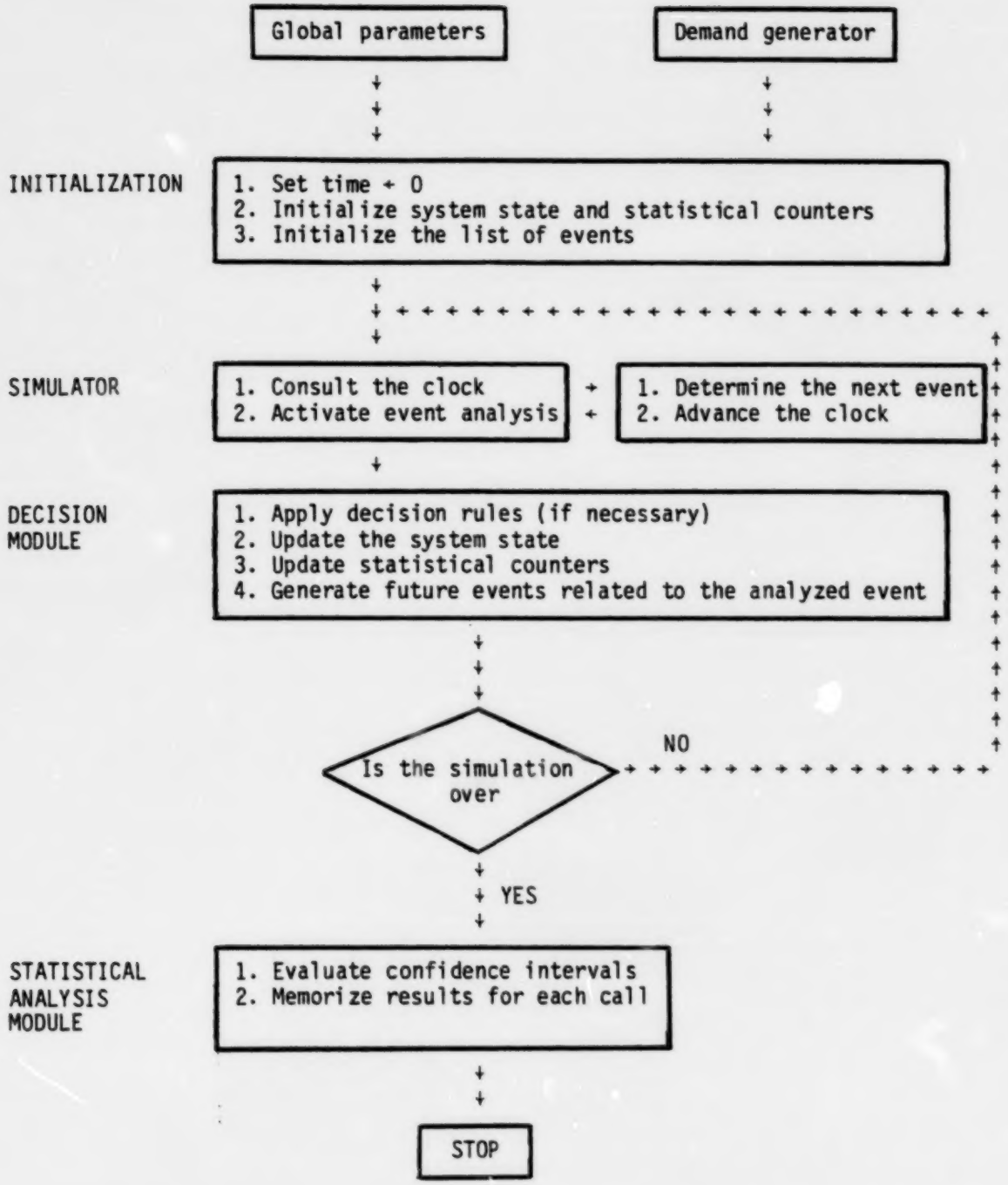


Figure 4. Simulation components.

- the status and location of every unit (for the dispatching and relocation of units);
- the origin of current calls (for the treatment of queues);
- the status of the waiting points (for relocation of units if any).

An event is a phenomenon occurring at a specific moment in time and unlatching an analytical process to evaluate its impact. This impact can be

- to modify the state of the system;
- to prompt a decision;
- to induce new events, i.e. its sole occurrence makes certain the existence of future events.

5.2 Simulation system components

Running a simulation requires several steps and a set of modules allows their execution. Figure 4 illustrates these. These steps are:

1) Fix the global parameters of the simulation.
2) Create a file of emergency calls (demand generator).
3) Select a strategy together with the parameters describing it.
4) Execute the simulation program itself. A file recording the results is created (one record per simulated call).
5) Complete the statistical analysis. The statistical module also includes procedures to determine the length of the simulation as a function of the confidence level required.

We now briefly describe each component.

5.3 Global parameters

They completely specify the situation that will be simulated.

1. General

   They include the type of the simulation (finite or infinite horizon) and a set of parameters that determine its length, the initial conditions and the arrival rate(s) of calls.

2. Territory representation

   The region to be covered is divided into sectors, and the demand is assumed to originate from their centroid. In Montreal a sector corresponds to a census sector. This representation is characterized by
   - a table of distances, i.e. the length of the shortest path for each pair of nodes, using the underlying network;
   - a predecessor table that lists the predecessor on the shortest path of every pair of nodes. This is necessary to locate ambulance in movement.

3. Speeds

   Two types of speeds are considered: one for emergency calls (those of highest priority) and one for other movements. The congestion on the network is dealt with by modifying arc lengths with multiplicative factors and then the use of a uniform speed.

4. Random variables

   The non deterministic elements of the system essentially relate to call occurrence and service time.

   - For call occurrence we have to specify:
     - the moment of occurrence
     - the source node
     - its priority (in Montreal there are three levels of priority);

- For service we have to specify:
    - the type of service: whether or not to take the patient to a hospital
    - service time: by definition, this includes the time from dispatching until the unit becomes available again. However, since the simulator calculates the times required for unit movements, only parameters related to the time spent at the site of the call and at the hospital have to be specified
    - the hospital to which the patient must be taken (if needed). Three variants are allowed: use the hospital nearest to the call; generate the hospital according to the global distribution of hospital's rates of occupancy; use the distribution conditional to the origin of the call.

5.4 Decision module

We include here the set of decisions that could be taken by the planner. These decisions are either taken in the planning stage, or in the operational stage (in real time).

Planning stage decisions:

- The number of units in service by time of day.
- Vehicles schedule: for each unit, specify the beginning and end of its workday. Breaks can also be included.
- Waiting points: when a unit is available it is located on a waiting point, although not necessarily always the same one. The number of these points and their location must be specified.

Operational stage

This describes the set of rules that enable decisions to be taken during the simulation concerning the dispatching of units, and their relocation.

- Dispatching

When a new call enters, one must decide:

1) How many ambulances must be dispatched?
2) Must this call be handled immediately or can it wait? If it has the highest priority then the answer is yes. Otherwise, depending on the strategy used, the call will be delayed if the total number of available resources is under a certain threshold.
3) When the call is handled, which unit should be sent ?

- Relocation

A relocation might occur whenever a unit becomes busy, or whenever a busy unit comes free again, but not necessarily at each instance of these events.

Hence one has to decide:

1) when to relocate the units according to the strategy adopted;
2) if a relocation is decided, which units are to be relocated and how is this done. In particular one should specify the path each unit should follow to reach its new waiting point. Recall that new events can occur during the relocation and hence the paths used are of some importance.

5.5 Statistical analysis module

The ultimate goal of the simulation is to evaluate a strategy or to compare strategies, using a simulated environment which must be as close as possible to reality. Statistical measures and good estimators for these measures are required in order to obtain an adequate confidence interval. But, the classical methods of statistical inference cannot be used in simulation and are not even adaptable, since simulations almost always generate correlated observations. Methods specially designed for simulation must be used. In particular special methods are required for the determination of initial conditions in the case of a finite horizon simulation, to obtain approximately independent observations, to determine the

minimal length of simulation required for a predefined level of confidence, and for the identification of the transient period before steady state is reached in case of infinite horizon (Law and Kelton, 1982; Bratley et al., 1987).

Some of these issues are illustrated in Figures 5 and 6 showing the cumulative mean response time (in minutes) with 50 units and an average workload ρ equal to 0.2 and 0.8 respectively, all other parameters being equal. These infinite horizon simulations illustrate how much the transient period and the length of simulation required can vary as ρ increases. The first 200 observations were identified as the transient period for ρ=0.2, whereas this number increased to 60,000 for ρ=0.8. Similarly, 300 000 observations were required for ρ=0.8 to attain a predetermined level of precision but only 7 000 for ρ=0.2.

The statistics used to measure the performances are

- the expected response time;
- the proportion of calls for which the response time is lower than a specific bound (five minutes for example);
- the working load of each unit (the percentage of time it is busy).

5.6 Conclusion

Simulation is expensive to develop, needs to be fed by a large amount of data and may require very long running times. Nevertheless, the EMS environment with its complexity and where saved lives may be at stake, can greatly benefit from such an evaluation tool. New strategies can be tested without taking any risks. Another possible application is to use interactive simulation as a training device (e.g. for new dispatchers). It can also be used to test various scenarios for situations that occur very infrequently but that have very serious implications (plane crash, congestion that completely change the time matrix e.g. when a bridge is unusable).

6. CONCLUSION

EMS operations are characterized by the presence of many processes that mutually interact so that the impact of a specific action on such a system is often difficult to evaluate. Operations research does not pretend to be a universal panacea, but it offers a solution framework that can produce "good" solutions to a large number of problems encountered in EMS. This paper has described a number of those solutions that can have a definite impact on the community. The authors have conducted other studies, among which one has permitted the reduction of the number of waiting points, that have helped Montreal EMS to achieve better efficiency. A number of problems though still remain intractable and present a great challenge. Analytical methods for relocation strategies and the presence of multiple objectives, mixed fleets and call priorities are certainly among these. But as our experience has demonstrated in Montreal, when the authorities are open to new solution approaches and when operations research practitioners are willing to try to adapt their models to reality, a close collaboration can be enriching to all parties.

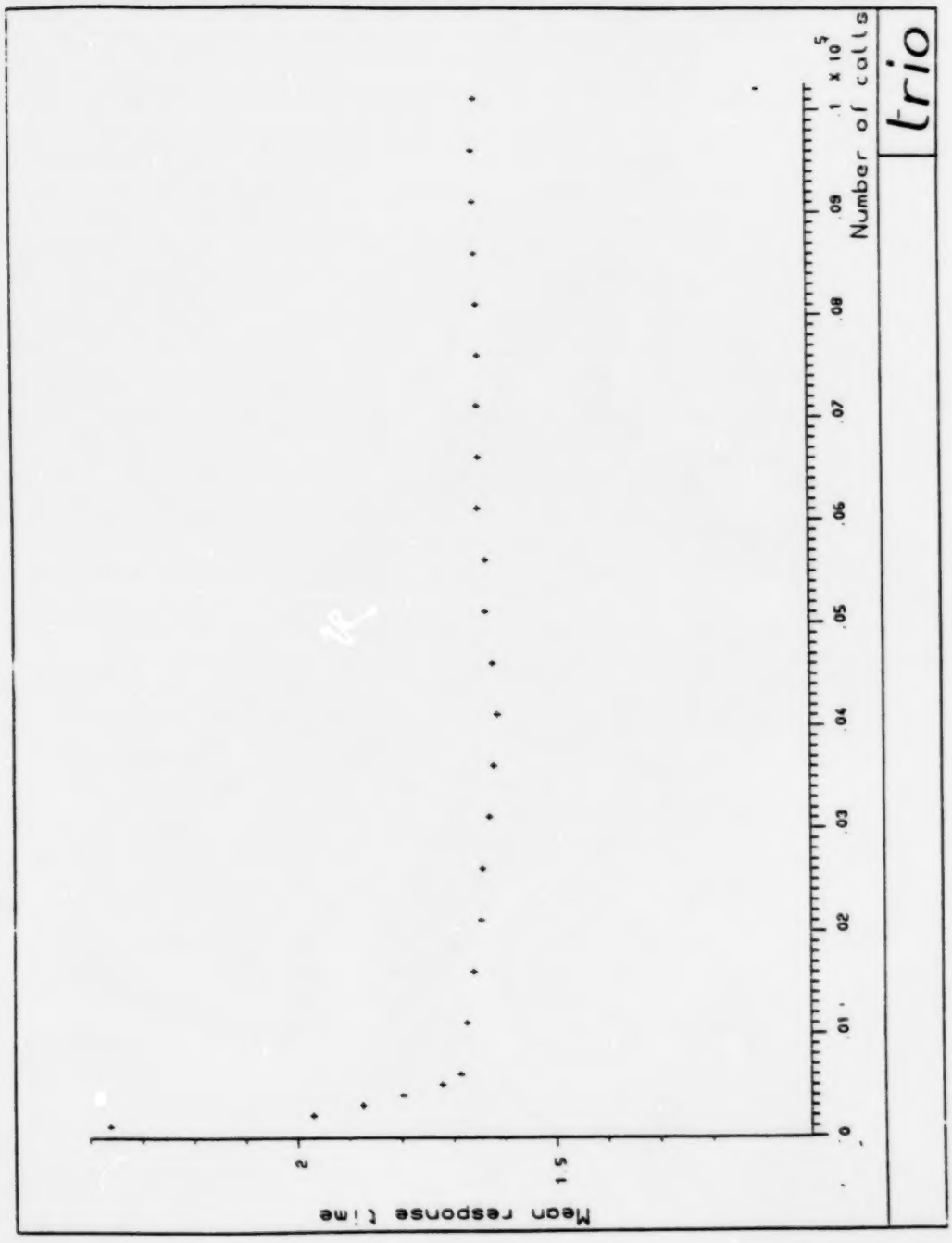


Figure 5: Simulation with n=50, ro=0.2

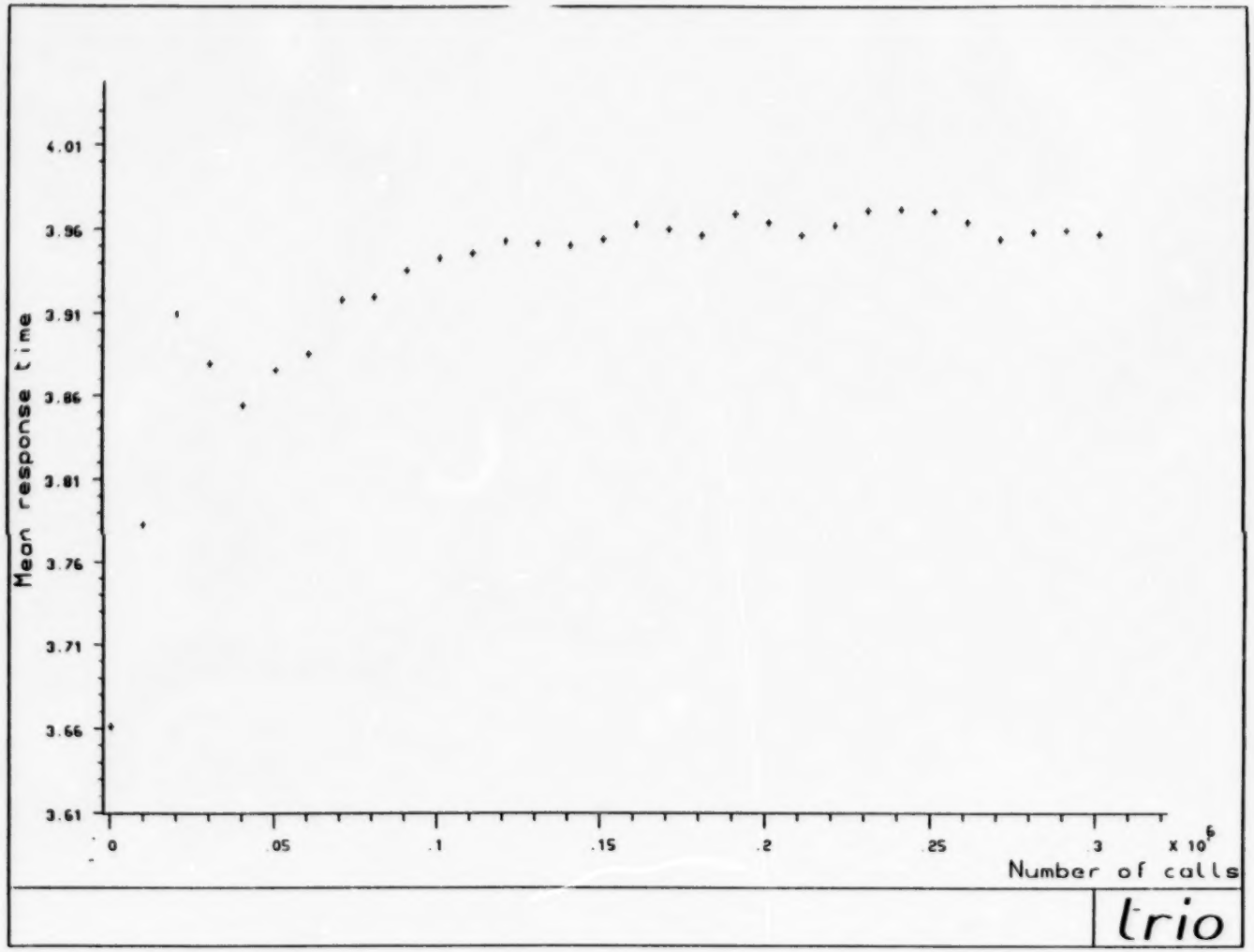


Figure 6: Simulation results for n=50, ro=0.8

References

(1) A. Babin, M. Dagenais, M. Florian, M. Gaudry, J. Guélat, P. Lestage, T.C. Liem, "TRIO - An open interactive graphic system for demand model estimation", Centre de recherche sur les transports, Université de Montréal, publication #441, 1987.

(2) O. Berman, "Dynamic repositioning of indistinguishable service units on transportation networks". Transportation Science, Vol 15, no 2, 115-136, 1981.

(3) O. Berman, "Repositioning of distinguishable urban service units on networks". Computers and Operations Research, Vol 8, 105-118, 1981.

(4) J.-Y. Blais, J.-M. Rousseau, "HASTUS: A model for the economic evaluation of drivers' collective agreements in transit companies". INFOR 20, 3-15, 1982.

(5) G.E.P. Box, G.M. Jenkins, "Time series analysis: forecasting and control". Holden-Day, San Francisco, 1976.

(6) P. Bratley, B. Fox, L. Schrage, "A guide to simulation". 2nd edition, Springer-Verlag, New York, 1987.

(7) G.M. Carter, J.M. Chaiken, E. Ignall, "Response areas for two emergency units". Operations Research, Vol 20, no 3, 571-594, 1972.

(8) J. Choquette, M.J.I. Gaudry, R. Roy, "Daily and hourly forecasts of the demand for EMS: a comparison of models". Centre de recherche sur les transports, Université de Montréal, (forthcoming) 1988.

(9) M.G. Dagenais, M.J.I. Gaudry, T.C. Liem, "Urban travel demand: the impact of Box-cox transformations with non spherical residual errors". Centre de recherche sur les transports, Université de Montréal, publication #358, 1985.

(10) T. Domenchich, D. McFadden, "Urban travel demand: a behavioural analysis", North-Holland Publishers, Amsterdam, 1975.

(11) M.J.I. Gaudry, "DRAG, un modèle de la demande routière des accidents et leur gravité, appliqué au Québec de 1956 à 1982". Centre de recherche sur les transports, Université de Montréal, publication #360, 1984.

(12) C. Gouriéroux, A. Montfort, "Cours de séries temporelles". Economica, Paris, 1983.

(13) J.P. Jarvis, "Optimization in stochastic service with distinguishable servers", TR-19-75, Innovative resource planning project in urban public safety systems". Operations Research Center, M.I.T., June 1975.

(14) J.P. Jarvis, "Optimal Assignments in a markovian queuing system". Comut & Ops. Res., Vol 8, 17-23, 1981.

(15) P. Kolesar, W.E. Walker, "An algorithm for the dynamic relocation of fire companies". Operations Research, Vol 22, no 2, 249-274, 1974.

(16) R.L. Larson, "A hypercube queuing model for facility location and redistricting in urban emergency services". Comput & Ops. Res., Vol 1, 67-95, 1974.

(17) R.L. Larson, "Approximating the performance of urban emergency service systems". Operations Research, Vol 23, no 5, 845-868, 1975.

(18) R.L. Larson, A.R. Odoni, "Urban operations research". Prentice-Hall Inc., Englewood Cliffs, N.J., 1981.

(19) P. Trudeau, J. Hodgson, "Localisation optimale de véhicules d'urgence". Centre de recherche sur les transports, Université de Montréal. À paraître comme publication, 1988.

(20) G.O. Wesolowsky, W.G. Truscott, "The multiperiod location-allocation problem with relocation of facilities". Management Science, Vol 22, no. 1, 57-65, 1975.